Stable Feature Selection for Biomarker Discovery

Zengyou He, Weichuan Yu
Laboratory for Bioinformatics and Computational Biology,
Department of Electronic and Computer Engineering,
The Hong Kong University of Science and Technology, Kowloon, Hong Kong, China.

January 7, 2010

Abstract

Feature selection techniques have been used as the workhorse in biomarker discovery applications for a long time. Surprisingly, the stability of feature selection with respect to sampling variations has long been under-considered. It is only until recently that this issue has received more and more attention. In this article, we review existing stable feature selection methods for biomarker discovery using a generic hierarchal framework. We have two objectives: (1) providing an overview on this new yet fast growing topic for a convenient reference; (2) categorizing existing methods under an expandable framework for future research and development.

Keywords: Feature selection; biomarker discovery; stability; machine learning

1 Introduction

Recent advances in genomics and proteomics enable the discovery of biomarkers for diagnosis and treatment of complex diseases at the molecular level [1]. A biomarker may be defined as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [2].

The discovery of biomarkers from high-throughput “omics” data is typically modeled as selecting the most discriminating features (or variables) for classification (e.g. discriminating healthy versus diseased, or different tumor stages) [3, 4]. In the language of statistics and machine learning, this is often referred to as feature selection. Feature selection has attracted strong research interest in the past several decades. For recent reviews of feature selection techniques used in bioinformatics, the reader is referred to [5, 6, 3, 7].

While many feature selection algorithms have been proposed, they do not necessarily identify the same candidate feature subsets if we repeat the biomarker discovery procedure [8]. Even for the same data, one may find many different subsets of features (either from the same feature selection method or from different feature selection methods) that can achieve the same or similar predictive accuracy [9, 10, 11]. In practice, high reproducibility of feature selection is equally important as high classification accuracy [12]. It is widely believed that a study that cannot be repeated has little value [13]. Consequently, the instability of feature selection results will reduce our confidence in discovered markers.

The stability issue in feature selection has received much attention recently. In this article, we shall review existing methods for stable feature selection in biomarker discovery applications, summarize them with an unified framework and provide a convenient reference for future research and development.

This article differs from existing review papers on feature selection in the following aspects:

- Compared to current feature selection reviews [5, 6, 3, 7], this review focuses only on those feature selection approaches that incorporate "stability" into the algorithmic design.
- This article mainly focuses on "methods" for finding reliable markers rather than "metrics" of measuring the stability of selected feature subsets [14], although we also list these metrics for completeness.

The remainder of the paper is organized as follows. In section 2, we discuss several sources that cause the instability of feature selection. In section 3, we summarize available stable feature selection algorithms and describe different classes of methods in detail. In section 4, we provide a list of stability measures and illustrate their definitions. We give a discussion in section 5. Finally, we conclude this paper in section 6.

2 Causes of Instability

There are mainly three sources of instability in biomarker discovery:

1. Algorithm design without considering stability: Classic feature selection methods aim at selecting a minimum subset of features to construct a classifier of the best predictive accuracy [8]. They often ignore "stability" in the algorithm design.
2. The existence of multiple sets of true markers: It is possible that there exist multiple sets of potential true markers in real data. On the one hand, when there are many highly correlated features, different ones may be selected under different settings [8]. On the other hand, even there are no redundant features, the existence of multiple non-correlated sets of real markers is also possible [15].
3. Small number of samples in high dimensional data: In the analysis of gene expression data and proteomics data, there are typically only hundreds of samples but thousands of features. It has been experimentally verified that the relatively small number of samples in high dimensional data is one of the main sources of the instability problem in feature selection [16, 17]. To understand the nature of the instability of selected feature subset, Ein-Dor et al [18] developed a new mathematical model and concluded that at least thousands of samples are needed to achieve stable feature selection.

Here we list three sources that can cause the instability of feature selection in biomarker discovery. We believe that there may be still other sources that can affect the stability of feature selection. The identification

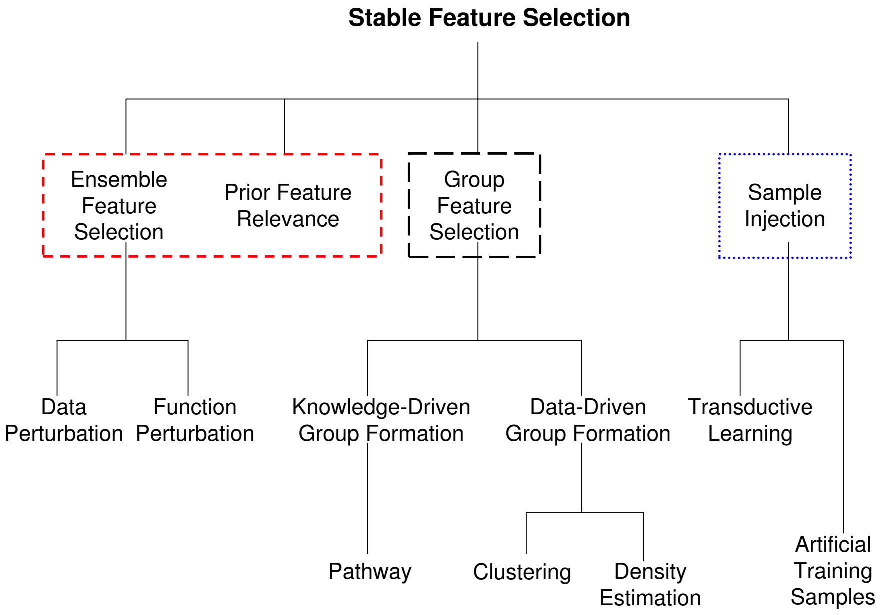

Figure 1: A hierarchical framework for stable feature selection methods.

of these sources is of primary importance for future research and development. On the one hand, knowing the reason enables us to better understand the problem. On the other hand, such knowledge will facilitate the design of new methods for stable biomarker discovery.

3 Existing Methods

To date, there are many methods available for stable feature selection. We wish to cover all existing methods in a systematic and expandable manner. Fig.1 illustrates our approach to summarizing different methods based on the way they treat different sources of instabilities. Briefly, the ensemble feature selection method and the method using prior feature relevance incorporate stability consideration into the algorithm design stage. To handle data with highly correlated features, the group feature selection approach treats feature cluster as the basic unit in the selection process to increase robustness. The sample injection method tries to increase the sample size to address the small-sample-size vs. large-feature-size issue. In the following sections, we will discuss each category in detail.

3.1 Ensemble Feature Selection

In statistics and machine learning, ensemble learning methods combine multiple learned models under the assumption that “two (or more) heads are better than one”. Typical ensemble learning methods such as bagging [19] and boosting [20] have been widely used in classification and regression. Ensemble feature selection techniques use a two-step procedure that is similar to ensemble classification and regression:

1. Creating a set of different feature selectors.

2. Aggregating the results of component selectors to generate the ensemble output.

The second step is typically modeled as a rank aggregation problem. Rank aggregation combines multiple rankings into a consensus ranking, which has been widely studied in the context of web search engines [21]. In most cases, the strategies rely on the following information:

- The ordinal rank associated with each feature.

- The score assigned to each feature.

To date, many rank aggregation approaches have been proposed and the reader is referred to [14] for a survey of popular aggregation methods used in bioinformatics.

Both theoretical and experimental results have suggested that the generation of a set of *diverse* component learners is one of the keys to the success of ensemble learning [22]. To construct diverse local learners, two strategies are widely used: data perturbation and function perturbation.

3.1.1 Data Perturbation

Data perturbation tries to run component learners with different sample subsets (e.g., Bagging [19], Boosting [20]) or in distinct feature subspaces (e.g., Random Subspace [23]). In ensemble feature selection with data perturbation, different samplings of the original data are generated to construct different feature selectors, as described in Fig.2. Several recent methods [24, 25, 26, 27] fall into this category. These methods can be further distinguished according to the sampling method, the component feature selection algorithm and the rank aggregation method (see Table 1).

The combination of data sampling and ensemble learning for feature selection is probably the most intuitive idea to handle selection instability with respect to sampling variation. The superiority of such strategy has been verified both experimentally [24, 27] and theoretically [25, 26].

Interestingly, all methods listed in Table 1 are based on the same aggregation scheme, i.e., linear combination. Note that it is also feasible to combine data perturbation with other complicated aggregation procedures such as those ones used in function perturbation (see next subsection).

3.1.2 Function Perturbation

Here we use function perturbation to refer to those ensemble feature selection methods in which the component learners are different from each other. The basic idea is to capitalize on the strengths of different algorithms to obtain robust feature subsets.

Function perturbation is different from data perturbation in two perspectives:

- It uses different feature selection algorithms rather than the same feature selection method.

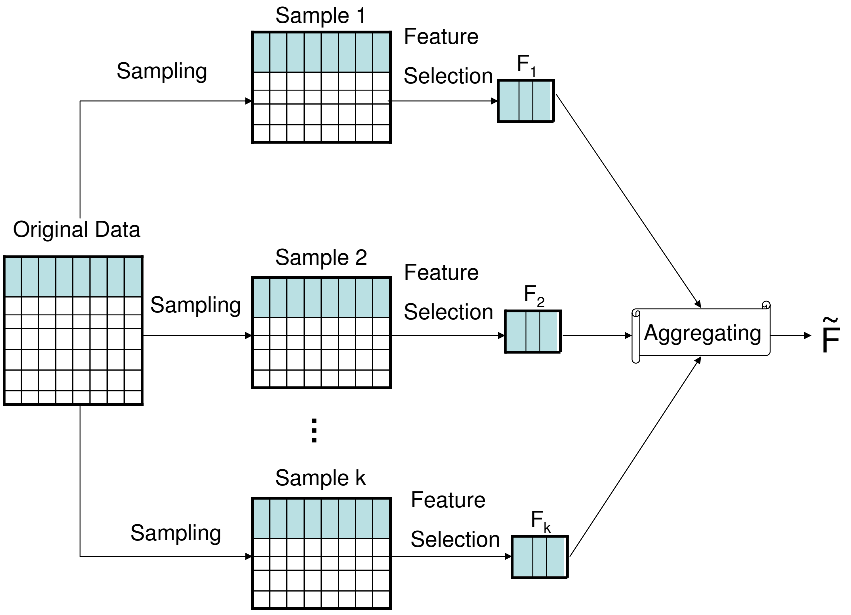

Figure 2: Ensemble feature selection using data perturbation. We first use different sub-samplings of training data to select features and then build a consensus output with a rank aggregation method.

- It typically conducts local feature selection on the original data (without sampling).

Existing ensemble feature selection methods in this category [30, 31, 32, 33] differ mainly in the aggregation procedure:

- The distance synthesis method is used in [30].
- The Markov chain based rank aggregation method [21] is utilized in [31].
- The linear combination method is used in [32].
- The concept of stacking [34] is applied to the aggregation of feature selection results in [33].

Compared to data perturbation, function perturbation is less flexible since the ensemble scale is limited by the number of available feature selection algorithms in the system. As a result, no more than four component feature selectors are used in the ensemble learning process [30, 31, 32, 33].

3.2 Feature Selection with Prior Feature Relevance

In most biomarker discovery applications, we typically assume that all features are equally relevant before the selection procedure. In practice, some prior knowledge may be available to bias the selection towards some features assumed to be more relevant [35, 36]. It has been shown that the use of prior knowledge on relevant features induces a large gain in stability with improved classification performance [35].

Table 1: Classification of data perturbation based ensemble feature selection methods. Here linear combination methods aggregate rankings using the (weighted) *min*, *max*, or *sum* operation. The filter method is a general feature selection strategy that attempts to rank features solely according to their relevance to target class.

Reference	Sampling Method	Feature Selector	Aggregation Method
Davis et al [24]	Random Subset	Filter Method	Linear Combination
Bach [25]	Boostrap	Lasso [28]	Linear Combination
Meinshausen and Buhlmann [26]	Random Subset	Randomized Lasso	Linear Combination
Abeel et al [27]	Random Subset	SVM-REF [29]	Linear Combination

Fig.3 shows that there are several methods for obtaining such kind of prior knowledge. One feasible method is to seek advices from domain experts or relevant publications. For instance, in gene expression data classification, one biologist may know or guess that some genes are likely to be more relevant [35].

Another more interesting method is to obtain such prior knowledge from relevant data sets using transfer learning [36]. Transfer learning focuses on extracting knowledge from source task and applying it to a different but related task [37]. In [36], those features that have been identified as markers from other data sets are considered to be more relevant in the new feature selection task.

Though the prior knowledge is helpful in improving the stability of feature selection, using such information deserves certain limitations since biomarker discovery aims at finding new features rather than known ones.

3.3 Group Feature Selection

One motivation for group feature selection is that groups of correlated features commonly exist in high-dimensional data, and such groups are resistant to the variations of training samples [16]. If each feature group is considered as a coherent single entity, potentially we may improve the selection stability.

Existing group feature selection algorithms follow the procedure described in Fig. 4. There are two key steps: group formation and group transformation.

Group formation is the process of identifying groups of associated features. There are typically two classes of methods for this purpose: knowledge-driven methods and data-driven methods. The knowledge-driven group formation method utilizes domain knowledge to facilitate the generation of feature groups. For example, genes normally function in co-regulated groups, making it feasible to search genes in the same pathway for group identification. In contrast, the data-driven group formation method finds feature clusters using only information contained in the input data.

Group transformation generates a single coherent representation for each feature group. The transformation method can range from simple approaches like feature value mean [38] to complicated methods such as principal component analysis [39].

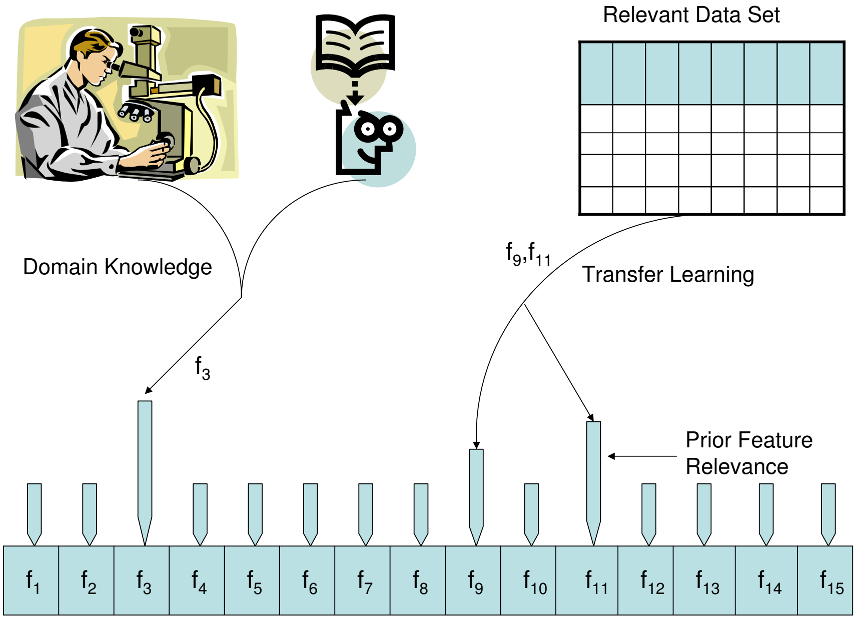

Figure 3: Stable feature selection using prior knowledge on features. The prior knowledge on relevant features are either obtained from domain experts or extracted from relevant data sets via transfer learning.

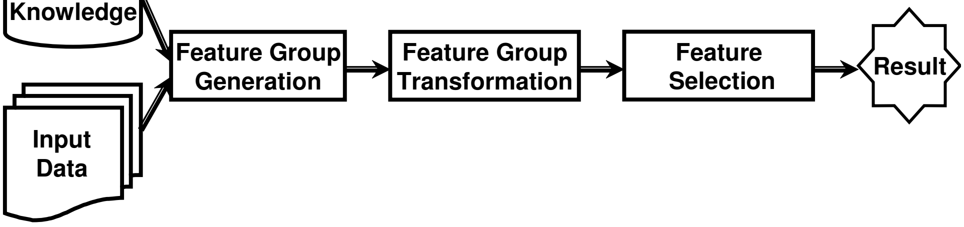

Figure 4: A generic group feature selection framework. In the first step, we identify feature groups using either knowledge-driven methods or data-driven approaches. In the second step, we transform each feature group into a single entity. Finally, we conduct feature selection in the transformed space.

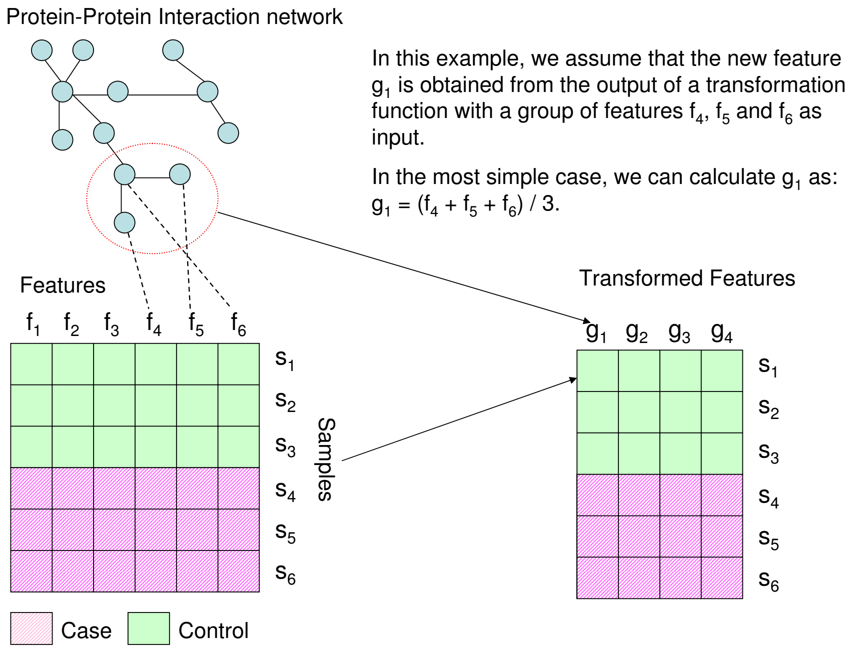
Protein-Protein Interaction network
In this example, we assume that the new feature g_1 is obtained from the output of a transformation function with a group of features f_4, f_5 and f_6 as input.
In the most simple case, we can calculate g_1 as: $g_1 = (f_4 + f_5 + f_6) / 3$.
Features
f_1 f_2 f_3 f_4 f_5 f_6
s_1 s_2 s_3 s_4 s_5 s_6
Samples
Transformed Features
g_1 g_2 g_3 g_4
s_1 s_2 s_3 s_4 s_5 s_6
Case
Control

Figure 5: An illustration of knowledge-driven group formation and feature transformation. The pathway information is used to guide the search of correlated features (genes or proteins). Each identified feature group is transformed into a new feature for further analysis.

In the following subsections, we will discuss existing group feature selection methods according to their group formation strategies.

3.3.1 Knowledge-Driven Group Formation

Recent advances in the construction of large protein networks make it feasible to find genes or proteins that have coherent expression patterns in the same pathway [1]. Using available protein-protein interaction (PPI) networks, a number of approaches have been proposed to incorporate the pathway information into the biomarker discovery procedure. As shown in Fig.5, the basic idea is to find a group of associated genes or proteins from the same pathway, and then transform this group into a new entity for subsequent feature selection and classification. It has been shown that such knowledge-based method is capable of achieving more reproducible feature selection and higher accuracy [40].

We can further distinguish available methods in this category according to their target data: gene expression data and proteomics data.

Before summarizing existing approaches on biomarker discovery using gene pathways, we would like to discuss a closely related problem: gene set significance testing. Testing the statistical significance of gene pathways or clusters has been extensively investigated. Well-known examples include the gene set enrichment

[1] For simplicity, this paper uses the terms “gene ontology”, “pathway”, and “gene set” interchangeably, although they may not be strictly equivalent.

analysis [41] and the maxmean approach [42]. The reader is referred to [43, 44] for comprehensive reviews of existing approaches on gene set analysis. Here we highlight the fact that gene set significance testing is different from pathway-guided biomarker discovery since they have different objectives. The objective of gene set significance testing is to find whether a given gene set satisfies the hypothesis, while biomarker discovery aims at searching for a small subset of genes that can distinguish cases from controls as accurately as possible. Their intrinsic connection is that we can utilize pathway significance assessment method as a filter method (a special type of feature selection technique that considers each entity individually) for ranking pathway markers.

Some methods have been proposed to identify markers not as individual genes but gene sets [38, 39, 40, 45, 46, 47, 48, 49, 50]. In Table 2, we provide a brief summary of existing biomarker identification methods that use gene pathway information as prior knowledge. These methods exploit different strategies for group generation and transformation. In group generation, we can use all genes in the pathway for a clear biological interpretation. Alternatively, we can search for a subset of genes so as to obtain one more discriminating group. To effectively represent each group, various summary statistics have been applied, ranging from mean to principal component analysis.

Table 2: Summary of gene pathway biomarker discovery methods. In the generation of gene groups, we either accept all genes in a given pathway or use heuristic search methods (such as greedy algorithm) to find a subset of discriminating genes. Here “No transformation” means that we use all the genes in the group to represent this gene set. GXNA (Gene eXpression Network Analysis) is a software package developed in [51] for identifying a subset of differentially expressed genes from a given pathway.

Reference	Group Generation	Group Transformation
Guo et al [38]	Use all genes	Mean and median
Rapaport et al [39]	Use all genes	Principal component analysis
Chuang et al [40]	Greedy search	Sum of z-scores
Tai and Pan [45]	Use all genes	No transformation
Lee et al [46]	Greedy search	Sum of z-scores
Hwang and Park [47]	Greedy search	Mean
Yousef et al [48]	GXNA [51]	No transformation
Chen and Wang [49]	Use all genes	Principal component analysis
Su et al [50]	Use all genes	Sum of log-likelihood ratio

Recently, such knowledge-driven approach has also been applied to proteomics data for biomarker discovery at the level of protein group [52, 53]. Compared to gene expression data, more research efforts toward this direction are desired in future research.

The pathway-guided group formation method has the advantage that new transformed features are biologically interpretable since the underlying disease process may be dependent on perturbations of different pathways. Thus, prediction models based on pathways may approximate the true disease process more closely than gene-based models [49]. Its main disadvantage is that we may group unrelated genes or proteins since the reliability of the predicted interactions in PPI network is still questionable [54].

3.3.2 Data-Driven Group Formation

Instead of relying on prior knowledge of biology, the data-driven group formation method identifies feature clusters using either cluster analysis [55, 56, 57, 58, 59, 60, 61, 62, 63] or density estimation [8, 16]. As summarized in Table 3, clustering-based methods utilize popular partition algorithms such as hierarchical clustering or k-means to generate feature groups. It should be noted that most existing clustering-based methods do not explicitly consider the stability of feature group. Alternatively, kernel density estimation is utilized in [8, 16] based on the observation that dense core regions are stable respect to samplings of dimensions.

Table 3: Summary of clustering-based group feature selection methods according to clustering algorithms and group transformation methods.

Reference	Clustering Methods	Group Transformation
Hastie et al [55]	Hierarchical clustering	Mean
Jornsten and Yu [56]	Integrated clustering and group selection	Mean
Au et al [57]	K-modes	One most discriminating feature
Ma et al [58]	K-means	A subset of most discriminating features
Ma and Haung [59]	Hierarchical clustering/K-means	A subset of most discriminating features
Yousel et al [60]	K-means	No transformation
Park et al [61]	Hierarchical clustering	Mean
Shin et al [62]	Hierarchical clustering	One most discriminating feature
Tang et al [63]	Fuzzy k-means	No transformation

There is another class of related methods assigning comparable coefficients to correlated, important variables. The “elastic net” [64] is a typical example in this category. We omit these methods in this survey since they didn’t explicitly identify feature groups.

The data-driven group feature selection method fully exploits the characteristics of target data so that it is widely applicable. One main drawback is that it is not easy to interpret and validate the selected feature group biologically. One possible remedy is to use a hybrid strategy that combines the data-driven method with the knowledge-driven method, as recently discussed in [65, 66].

3.4 Feature Selection with Sample Injection

In biomarker discovery applications, the number of features is typically larger than the sample size. This is one of the main sources of instability in feature selection. To increase the reproducibility of feature selection, one natural idea is to generate more samples. However, the generation of real sample data from patients and healthy people is usually expensive and time-consuming. With this practical limitation in mind, people begin to seek other alternative methods for the same purpose.

From the viewpoint of data analysis, there are two data augmentation strategies:

- Utilizing test data to increase the sample size in feature selection process, which can be modeled as a transductive learning problem [67].
- Generating some artificial training samples according to the distribution of available training data.

In the following sections, we will introduce each method in detail.

3.4.1 Method Using Transductive Learning

Different from inductive learning algorithms, the transductive learning algorithm is not required to produce a general hypothesis that can predict the label of any unobserved data [67]. As illustrated in Fig. 6, it is only required to predict the labels of a given test set of samples. In other words, we can use both training data and testing data in the learning procedure.

Transductive learning has been used to increase sample size in some recent papers [68, 69]. The main idea is to take advantage of the information embedded in the test data so that the role of test samples is changed from *passive* to *active*. That is, the unlabeled test samples are incorporated into the feature selection and classification process.

3.4.2 Method Using Artificial Training Samples

The idea is to generate a number of artificial training samples according to the distribution of given samples. Then, feature subsets can be assessed using both the generated data and the original data. In Fig.7, we provide an example to illustrate the effect of injected artificial samples on model selection.

There are many methods for generating artificial training samples. For instance, we can first pick one training sample x_i randomly and then generate a point z from standard normal distribution. Finally, we generate the new artificial point as: $y = x_i + hz$, where h is a constant.

There are two ways in which the artificial points participate in feature selection. One method is to treat the injected points as the original samples in the training process [70]. Another method is to use the injected points only in the evaluation stage [71].

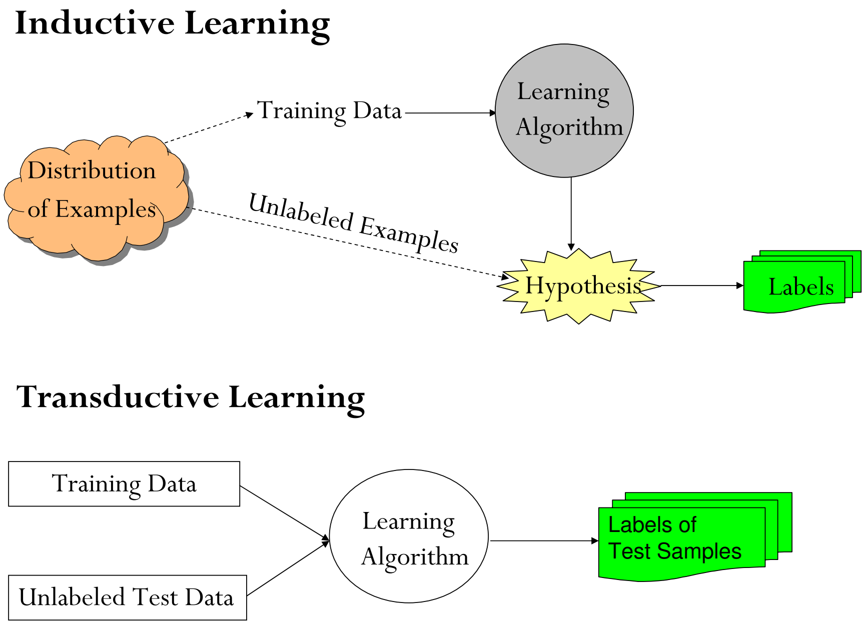

Figure 6: An illustration of inductive learning and transductive learning. Intuitively, we can consider inductive learning as "an education system for all-round development" while transductive learning is an "examination-oriented education system".

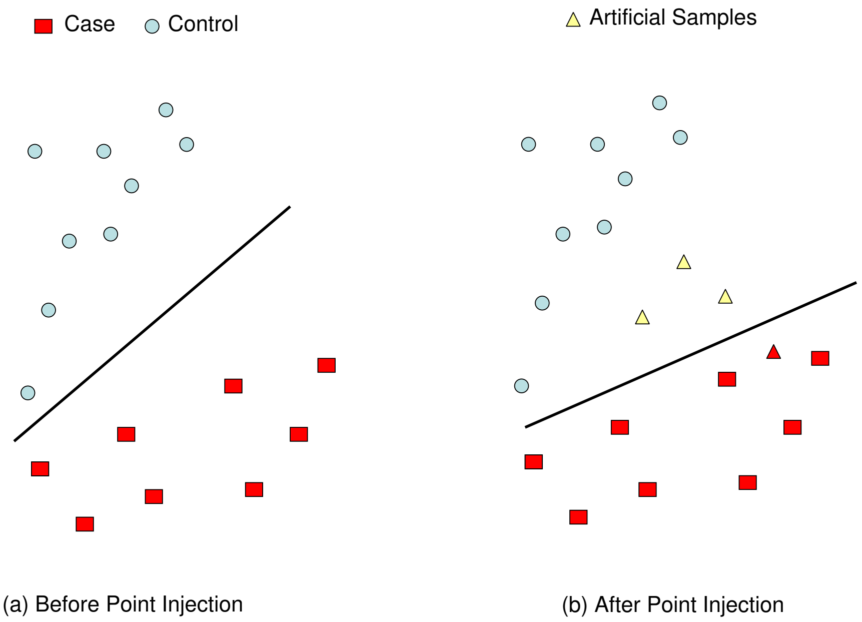

Figure 7: The effect of artificial training samples on model selection. The separating hyperplane obtained from the data set with injected samples is different from that of the original samples.

4 Stability Measure

In stable feature selection, one important issue is how to measure the "stability" of feature selection algorithms, i.e., how to qualify the selection sensitivity to variations in the training set. The stability measure can be used in different contexts. On the one hand, it is indispensable for evaluating different algorithms in performance comparison. On the other hand, it can be used for internal validation in feature selection algorithms that take into account stability.

Noticing that there is already a nice review paper [14] on the stability of ranked gene lists, here we would like to provide a more comprehensive list that includes evaluation methods from different domains.

Measuring stability requires a similarity measure for feature selection results. This depends on the representations used by different algorithms. Formally, let training examples be described by a vector of features $F = (f_1, f_2, ..., f_m)$, then there are three types of representation methods [72]:

- A subset of features: $S = \{s_1, s_2, .., s_k\}$, $s_i \in \{f_1, f_2, ..., f_m\}$.
- A ranking vector: $R = (r_1, r_2, .., r_m)$, $1 \leq r_i \leq m$.
- A weighting-score vector: $W = (w_1, w_2, .., w_m)$, $w_i \in R^+$.

In general, we are interested in stability measures that take more than two subsets (or rankings) into account. In this review, we use measures defined on two subsets (or rankings) for the sake of notation simplification. As pointed out in [14], there are essentially two approaches for generalizing the definition. One approach is to summarize pairwise stability measures through averaging. Another approach is to consider all subsets (or rankings) simultaneously in the specification of stability measure.

In the following, we will summarize available stability measures according to the representation of feature selection results.

4.1 Feature Subset

There is a wide variety of similarity measures available for the comparison of sets. Table 4 summarizes available stability measures. One may find that most of these measures are defined using the physical properties of two sets, e.g., the ratio of the intersection to the union. One exception is the "percentage of overlapping features related" [13], which incorporates additional feature correlation information into the measure definition. This is definitely plausible for biomarker discovery applications since there are always some highly correlated features in the "omics" data.

MS1: The relative Hamming distance between the masks corresponding to two subsets is used to measure the stability [73].

MS2: The Tanimoto distance metric measures the amount of overlap between two sets of arbitrary cardinality. It takes values in [0,1], with 0 meaning no overlap between the two sets and 1 meaning two sets are identical. In fact, this measure is equivalent to the Jaccard's index: $\frac{|S \cap S'|}{|S \cup S'|}$.

Table 4: A list of stability measures when the feature selection algorithm produces a subset of features as output. Here S and $S^{'}$ are two subsets of features.

Index	Description	Formula
MS1 [73]	Relative Hamming distance	$1-\frac{\|S\backslash S^{'}\|+\|S^{'}\backslash S\|}{m}$
MS2 [72, 74, 75]	Tanimoto distance/Jaccard's index	$1-\frac{\|S\|+\|S^{'}\|-2\|S\cap S^{'}\|}{\|S\|+\|S^{'}\|-\|S\cap S^{'}\|}$
MS3 [8, 11, 16]	Dice-Sorensen's index	$\frac{2\|S\cap S^{'}\|}{\|S\|+\|S^{'}\|}$
MS4 [11]	Ochiai's index	$\frac{2\|S\cap S^{'}\|}{\sqrt{\|S\|\|S^{'}\|}}$
MS5 [76]	Percentage of overlapping features	$\frac{\|S\cap S^{'}\|}{\|S\|}, \frac{\|S\cap S^{'}\|}{\|S^{'}\|}$
MS6 [13]	Percentage of overlapping features related	$\frac{\|S\cap S^{'}\|+c_{12}}{\|S\|}, \frac{\|S\cap S^{'}\|+c_{21}}{\|S^{'}\|}$
MS7 [77]	Kuncheva's stability measure	$\frac{\|S\cap S^{'}\|m-c^2}{c(m-c)}$
MS8 [78]	Consistency	$\frac{1}{\|S\cup S^{'}\|}\sum\limits_{f\in S\cup S^{'}}\frac{freq(f)-1}{m-1}$
MS9 [78]	Weighted consistency	$\sum\limits_{f\in S\cup S^{'}}(\frac{freq(f)}{\|S\|+\|S^{'}\|}\cdot\frac{freq(f)-1}{m-1})$
MS10 [24]	Length adjusted stability	$max\{0, \sum\limits_{f\in S\cup S^{'}}(\frac{freq(f)}{2\|S\cup S^{'}\|}-\alpha\frac{\|S\|+\|S^{'}\|}{2m})\}$

MS3: The Dice-Sorensen's index is the harmonic mean of $\frac{|S\cap S^{'}|}{|S|}$ and $\frac{|S\cap S^{'}|}{|S^{'}|}$.

MS4: The Ochiai's index is the geometric mean of $\frac{|S\cap S^{'}|}{|S|}$ and $\frac{|S\cap S^{'}|}{|S^{'}|}$. It has been shown that the performance of the Ochiai's index is similar with that of Jaccard's index and Dice-Sorensen's index [11].

MS5: This measure is originally named as: "Percentage of Overlapping Genes (POG)" in the context of gene expression data analysis.

MS6: It is an extension of POG, which incorporates highly correlated features between two sets into the stability evaluation. In the formula, c_{12} (or c_{21}) denotes the number of features in S (or $S^{'}$) that are not shared but are significantly positively correlated with at least one feature in $S^{'}$ (or S). The normalized form of this measure is also presented in [13].

MS7: This stability measure assumes that S and $S^{'}$ have the same size (cardinality), i.e., $|S| = |S^{'}| = c$.

MS8 and MS9: In both definitions, $freq(f)$ denotes the number of occurrences (frequency) of feature f in $S \cup S^{'}$. It has been proved that both measures take values in [0,1]. The (weighted) consistency value is 1 if two sets are identical and 0 if they are disjoint.

MS10: In the formula, α is one user-specified parameter and is set to 10 [24]. Note that $(|S| + |S^{'}|)/2$ corresponds to the median required in [24] since there are only two sets in our formulation.

4.2 Ranking List

The problem of comparing ranking lists is widely studied in different contexts such as voting theory and web document retrieval. Table 5 shows some distance measures for two ranking lists. One typical example is

MR2, in which the Spearman's correlation is adapted to place more weights on those top ranked features since these features are more important than irrelevant features in the stability evaluation.

Table 5: A list of stability measures when the feature selection algorithm generates a ranking list as output. Here R and $R^{'}$ are two different ranking lists.

Index	Description	Formula
MR1 [72]	Spearman's rank correlation coefficient	$1-6\sum\limits_{i=1}^{m}\frac{(r_i-r_i^{'})^2}{m(m^2-1)}$
MR2 [12]	Canberra distance	$\sum\limits_{i=1}^{m}\frac{\|min\{r_i,k+1\}-min\{r_i^{'},k+1\}\|}{min\{r_i,k+1\}+min\{r_i^{'},k+1\}}$
MR3 [79]	Overlap score	$\sum\limits_{i=1}^{m}e^{-\alpha i}\|\{r_j\|j<i\}\cap\{r_j^{'}\|j<i\}\|$

MR1: The Spearman's rank correlation coefficient takes values in [-1,1], with 1 meaning that the two ranking lists are identical and a value of -1 meaning that they have exactly inverse orders.

MR2: The Canberra distance is a weighted version of Spearmans footrule distance [12], i.e., $\sum\limits_{i=1}^{m}\frac{|r_i-r_i^{'}|}{r_i+r_i^{'}}$. Since the most important features are usually located at the top of the ranking list [12], the distance calculation in the table only considers top k ranked features.

MR3: The overlap score is originally proposed in [79] and here we follow [14] to reformulate it with the assumption that only top ranked features are important. In the formula, α is a user-specified parameter to control the decreasing rate.

4.3 Weighting-Score Vector

The computational issue of combinatorial search for feature subset can to some extent be alleviated by using a feature weighting strategy [80]. Allowing feature weights to take real-valued numbers instead of binary ones enables us to use well-established optimization techniques in algorithmic development. For instance, the RELIEF algorithm [81] is one representative of such kind of methods, which generates a weighting score vector as output.

The weighting score vector is seldom used in defining stability measure. Table 6 lists one stability measure MW1 [72]. The Pearson's correlation coefficient ranges from -1 to 1. A value of 1 indicates a perfect positive correlation, a value of 0 means that there is no correlation, while a value of -1 means that they are anti-correlated. In the formula, u_W and $u_{W'}$ are the means of weight scores of W and $W^{'}$, respectively.

5 Discussions

We summarize three sources of instability for feature selection in section 2. Among these sources, probably the small number of samples in high dimensional feature space is the most difficult one in biomarker discovery.

Table 6: The Pearson's correlation coefficient measure. Here W and $W^{'}$ are two different weighting score vectors.

Index	Description	Formula
MW1 [72]	Pearson's correlation coefficient	$\frac{\sum_i (w_i - u_W)(w_i^{'} - u_{W^{'}})}{\sqrt{\sum_i (w_i - u_W)^2 \sum_i (w_i^{'} - u_{W^{'}})^2}}$

Besides feature selection, other data analysis tasks also face the same challenges. Research progresses in related fields will facilitate the development of effective stable feature selection methods as well.

Group feature selection is the most extensively studied method among existing stable feature selection approaches. This is because there are many correlated features in high dimensional space. However, such feature grouping strategy can only partially alleviate selection instability since we still need to face the reproducibility issue in the transformed space. In this regard, ensemble feature selection is probably more promising to provide a general-purpose solution. One immediate hybrid strategy is to combine group feature selection with ensemble feature selection, i.e., first perform feature grouping and then use ensemble feature selection in the new feature space.

The group feature selection strategy is only helpful when multiple sets of true markers are generated due to the existence of redundant features. However, it is also possible that multiple sets of true markers share no correlated features. The feature selection problem in this case is much harder than finding a minimal optimal feature set for classification [82]. To our knowledge, there is still no available method and measure that aim at handling stability issues in this context. The general problem is open and needs more research efforts.

With respect to stability index, most available measures are defined over feature subsets since the feature subset can be obtained from rankings or scores (but not vice-versa). The major problem is that there is still no consensus on the best stability measure. Therefore, a comprehensive comparison study on existing stability measures should be conducted in future research.

In fact, the biomarker discovery process involves many procedures. Here we only discuss feature selection techniques for stable biomarker identification. The development of biomarker classifier is also very important. The readers are referred to a recent review [83] for research progress towards this direction.

Finally, we would like to raise the following questions in the pursuit of stable biomarker discovery methods for future research:

- How to directly measure the stability of feature(s) without sampling training data?
- Can we propose new methods that are capable of explicitly controlling the stability of reported feature subset?
- Are there other special requirements for biomarker discovery rather than stability?

6 Conclusions

To discover reproducible markers from "omics" data, the stability issue of feature selection has received much attention recently. This review summarizes existing stable feature selection methods and stability measures. Stable feature selection is a very important research problem, from both theoretical perspective and practical aspect. More research efforts should be devoted to this challenging topic.

Acknowledgments

This work was supported with the general research fund 621707 from the Hong Kong Research Grant Council, a research proposal competition award RPC07/08.EG25 and a postdoctoral fellowship from the Hong Kong University of Science and Technology.

References

[1] P. Srinivas, M. Verma, Y. Zhao, and S. Srivastava, "Proteomics for cancer biomarker discovery," *Clinical Chemistry*, vol. 48, no. 8, pp. 1160–1169, 2002.

[2] Biomarkers Definitions Working Group, "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework," *Clinical Pharmacology & Therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[3] M. Hilario and A. Kalousis, "Approaches to dimensionality reduction in proteomic biomarker studies," *Briefings in Bioinformatics*, vol. 9, no. 2, pp. 102–118, 2008.

[4] F. Azuaje, Y. Devaux, and D. Wagner, "Computational biology for cardiovascular biomarker discovery," *Briefings in Bioinformatics*, vol. 10, no. 4, pp. 367–377, 2009.

[5] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[6] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Briefings in Bioinformatics*, vol. 9, no. 5, pp. 392–403, 2008.

[7] B. Duval and J. Hao, "Advances in metaheuristics for gene selection and classification of microarray data," *Briefings in Bioinformatics*, 2009.

[8] L. Yu, C. Ding, and S. Loscalzo, "Stable feature selection via dense feature groups," in *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'08)*, 2008, pp. 803–811.

[9] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?" *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005.

[10] S. Michiels, S. Koscielny, and C. Hill, “Prediction of cancer outcome with microarrays: a multiple random validation strategy,” *Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

[11] M. Zucknick, S. Richardson, and E. A. Stronach, “Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods,” *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, p. 7, 2008.

[12] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello, “Algebraic stability indicators for ranked lists in molecular profiling,” *Bioinformatics*, vol. 24, no. 2, pp. 258–264, 2008.

[13] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, Q. Liu, J. Wang, D. Wang, C. Wang, and Z. Guo, “Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes,” *Bioinformatics*, vol. 25, no. 13, pp. 1662–1668, 2009.

[14] A. L. Boulesteix and M. Slawski, “Stability and aggregation of ranked gene lists,” *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 556–568, 2009.

[15] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, H. Xiao, D. Wang, D. Yang, X. Gong, J. Zhu, Y. Li, and X. Li, “Apparently low reproducibility of true differential expression discoveries in microarray studies,” *Bioinformatics*, vol. 24, no. 18, pp. 2057–2063, 2008.

[16] S. Loscalzo, L. Yu, and C. Ding, “Consensus group stable feature selection,” in *Proceeding of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (KDD’09)*, 2009, pp. 567–575.

[17] S.-Y. Kim, “Effects of sample size on robustness and prediction accuracy of a prognostic gene signature,” *BMC Bioinformatics*, vol. 10, p. 147, 2009.

[18] L. Ein-Dor, O. Zuk, and E. Domany, “Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 15, pp. 5923–5928, 2006.

[19] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[20] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[21] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, “Rank aggregation methods for the web,” in *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 613–622.

[22] G. Brown, J. Wyatt, R. Harris, and X. Yao, “Diversity creation methods: A survey and categorisation,” *Information Fusion*, vol. 6, no. 1, pp. 5–20, 2005.

[23] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[24] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kuffner, and R. Zimmer, "Reliable gene signatures for microarray classification: assessment of stability and performance," *Bioinformatics*, vol. 22, no. 19, pp. 2356–2363, 2006.

[25] F. R. Bach, "Bolasso: Model consistent lasso estimation through the bootstrap," in *Proceedings of the 25th Annual International Conference on Machine Learning (ICML'08)*, 2008, pp. 33–40.

[26] N. Meinshausen and P. Buhlmann, "Stability selection," *Preprint, arXiv*, 2008.

[27] T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, in press.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 58, no. 1, pp. 267–288, 1996.

[29] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[30] Y. Yang, Y. Xiao, and M. Segal, "Identifying differentially expressed genes from microarray experiments via statistic synthesis," *Bioinformatics*, vol. 21, no. 7, pp. 1084–1093, 2005.

[31] J. Dutkowski and A. Gambin, "On consensus biomarker selection," *BMC Bioinformatics*, vol. 8, no. Suppl 5, p. S5, 2007.

[32] N. C. Tan, W. G. Fisher, K. P. Rosenblatt, and H. R. Garner, "Application of multiple statistical tests to enhance mass spectrometry-based biomarker discovery," *BMC Bioinformatics*, vol. 10, p. 144, 2009.

[33] M. Netzer, G. Millonig, M. Osl, B. Pfeifer, S. Praun, J. Villinger, W. Vogel, and C. Baumgartner, "A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry," *Bioinformatics*, vol. 25, no. 7, pp. 941–947, 2009.

[34] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[35] T. Helleputte and P. Dupont, "Partially supervised feature selection with regularized linear models," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, 2009, pp. 409–416.

[36] ——, "Feature selection by transfer learning with linear regularized models," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, 2009, pp. 533–547.

[37] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Data and Knowledge Engineering*, in press.

[38] Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. Topol, Q. Wang, and S.Rao, “Towards precise classification of cancers based on robust gene functional expression profiles,” *BMC Bioinformatics*, vol. 6, p. 58, 2005.

[39] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J. Vert, “Classification of microarray data using gene networks,” *BMC Bioinformatics*, vol. 8, p. 35, 2007.

[40] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, “Network-based classification of breast cancer metastasis,” *Molecular Systems Biology*, vol. 3, p. 140, 2007.

[41] A. Subramanian, P. Tamayoa, V. K. Moothaa, S. Mukherjeed, B. L. Eberta, M. A. Gillettea, A. Paulovichg, S. L. Pomeroyh, T. R. Goluba, E. S. Landera, and J. P. Mesirova, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[42] B. Efron and R. Tibshirani, “On testing the significance of sets of genes,” *Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129, 2007.

[43] D. Nam and S. Y. Kim, “Gene-set approach for expression pattern analysis,” *Briefings in Bioinformatics*, vol. 9, no. 3, pp. 189–197, 2008.

[44] I. Dinu, J. Potter, T. Mueller, Q. Liu, A. Adewale, G. Jhangri, G. Einecke, K. Famulski, P. Halloran, and Y. Yasui, “Gene-set analysis and reduction,” *Briefings in Bioinformatics*, vol. 10, no. 1, pp. 24–34, 2009.

[45] F. Tai and W. Pan, “Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms,” *Bioinformatics*, vol. 23, no. 14, pp. 1775–1782, 2007.

[46] E. Lee, H. Chuang, J. Kim, T. Ideker, and D. Lee, “Inferring pathway activity toward precise disease classification,” *PLoS Computational Biology*, vol. 4, no. 11, p. e1000217, 2008.

[47] T. Hwang and T. Park, “Identification of differentially expressed subnetworks based on multivariate ANOVA,” *BMC Bioinformatics*, vol. 10, p. 128, 2009.

[48] M. Yousef, M. Ketany, L. Manevitz, L. Showe, and M. K. Showe, “Classification and biomarker identification using gene network modules and support vector machines,” *BMC Bioinformatics*, vol. 10, p. 337, 2009.

[49] X. Chen and L. Wang, “Integrating biological knowledge with gene expression profiles for survival prediction of cancer,” *Journal of Computational Biology*, vol. 16, no. 2, pp. 265–278, 2009.

[50] J. Su, B.-J. Yoon, and E. R. Dougherty, “Accurate and reliable cancer classification based on probabilistic inference of pathway activity,” *PLoS One*, 2010.

[51] S. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes, “Gene expression network analysis and applications to immunology,” *Bioinformatics*, vol. 23, no. 7, pp. 850–858, 2007.

[52] G. Jin, X. Zhou, H. Wang, H. Zhao, K. Cui, X. Zhang, L. Chen, S. Hazen, K.Li, and S. Wong, “The knowledge-integrated network biomarkers discovery for major adverse cardiac events,” *Journal of Proteome Research*, vol. 7, no. 9, pp. 4013–4021, 2008.

[53] K. Lawlor, A. Nazarian, L. Lacomis, P. Tempst, and J. Villanueva, “Pathway-based biomarker search by high-throughput proteomics profiling of secretomes,” *Journal of Proteome Research*, vol. 8, no. 3, pp. 1489–1503, 2009.

[54] C. von Mering, R. Krause, B. Snel, M. Cornell, S. Oliver, S. Fields, and P. Bork, “Comparative assessment of large-scale data sets of protein-protein interactions,” *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[55] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown, “Supervised harvesting of expression trees,” *Genome Biology*, vol. 2, no. 1, pp. research0003.1–0003.12, 2001.

[56] R. Jornsten and B. Yu, “Simultaneous gene clustering and subset selection for sample classification via mdl,” *Bioinformatics*, vol. 19, no. 9, pp. 1100–1109, 2003.

[57] W. Au, K. Chan, A. Wong, and Y. Wang, “Attribute clustering for grouping, selection, and classification of gene expression data,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 2, pp. 83–101, 2005.

[58] S. Ma, X. Song, and J. Huang, “Supervised group lasso with applications to microarray data analysis,” *BMC Bioinformatics*, vol. 8, p. 60, 2007.

[59] S. Ma and J. Huang, “Clustering threshold gradient descent regularization: with applications to microarray studies,” *Bioinformatics*, vol. 23, no. 4, pp. 466–472, 2007.

[60] M. Yousef, S. Jung, L. Showe, and M. Showe, “Recursive Cluster Elimination(RCE) for classification and feature selection from gene expression data,” *BMC Bioinformatics*, vol. 8, p. 144, 2007.

[61] M. Park, T. Hastie, and R. Tibshirani, “Averaged gene expressions for regression,” *Biostatistics*, vol. 8, no. 2, pp. 212–227, 2007.

[62] H. Shin, B. Sheu, M. Joseph, and M. K. Markey, “Guilt-by-association feature selection: Identifying biomarkers from proteomic profiles,” *Journal of Biomedical Informatics*, vol. 41, no. 1, pp. 124–136, 2008.

[63] Y. Tang, Y. Zhang, Z. Huang, X. Hu, and Y. Zhao, “Recursive fuzzy granulation for gene subsets extraction and cancer classification,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 6, pp. 723–730, 2008.

[64] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B(Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[65] M. Shi and S. Ma, “Identifying subset of genes that have influential impacts on cancer progression: a new approach to analyze cancer microarray data,” *Functional & Integrative Genomics*, vol. 8, no. 4, pp. 361–373, 2008.

[66] S. Ma, J. Huang, and S. Shen, “Identification of cancer-associated gene clusters and genes via clustering penalization,” *Statistics and Its Interface*, vol. 2, pp. 1–11, 2009.

[67] V. Vapnik, *Statistical learning theory*. Wiley Interscience, 1998.

[68] M. Zhu and A. M. Martinez, “Using the information embedded in the testing sample to break the limits caused by the small sample size in microarray-based classification,” *BMC Bioinformatics*, vol. 9, p. 280, 2008.

[69] T. Hwang, H. Sicotte, Z. Tian, B. Wu, J. Kocher, D. Wigle, V. Kumar, and R. Kuang, “Robust and efficient identification of biomarkers by classifying features on graphs,” *Bioinformatics*, vol. 24, no. 18, pp. 2023–2029, 2008.

[70] S. Kim, E. R. Dougherty, J. Barrera, Y. Chen, M. Bittner, and J. Trent, “Strong feature sets from small samples,” *Journal of Computational Biology*, vol. 9, no. 1, pp. 127–146, 2002.

[71] D. Huang and T. Chow, “Effective gene selection method with small sample sets using gradient-based and point injection techniques,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 467–475, 2007.

[72] A. Kalousis, J. Prados, and M. Hilario, “Stability of feature selection algorithms: a study on high-dimensional spaces,” *Knowledge and Information Systems*, vol. 12, no. 1, pp. 95–116, 2007.

[73] K. Dunne, P. Cunningham, and F. Azuaje, “Solutions to instability problems with sequential wrapper-based approaches to feature selection,” Department of Computer Science, Trinity College, Dublin, Tech. Rep. TCD-CS-2002-28, 2002.

[74] G. Stolovitzky, “Gene selection in microarray data: the elephant, the blind men and our algorithms,” *Current Opinion in Structural biology*, vol. 13, no. 3, pp. 370–376, 2003.

[75] R. Nilsson, J. Bjorkegren, and J. Tegner, “On reliable discovery of molecular signatures,” *BMC Bioinformatics*, vol. 10, p. 38, 2009.

[76] L. Shi, et al, “Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential,” *BMC Bioinformatics*, vol. 6, no. Suppl 2, p. S12, 2005.

[77] L. Kuncheva, “A stability index for feature selection,” in *Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*, 2007, pp. 390–395.

[78] P. Somol and J. Novovicová, “Evaluating the stability of feature selectors that optimize feature subset cardinality,” in *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, 2008, pp. 966–976.

[79] X. Yang, S. Bentink, S. Scheid, and R. Spang, “Similarities of ordered gene lists,” *Journal of Bioinformatics and Computational Biology*, vol. 4, no. 3, pp. 693–708, 2006.

[80] Y. Sun, “Iterative relief for feature weighting: Algorithms, theories, and applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1035–1051, 2007.

[81] K. Kira and L. Rendell, “A practical approach to feature selection,” in *Proceedings of the 9th International Workshop on Machine Learning*, 1992, pp. 249–256.

[82] R. Nilsson, J. Pena, J. Bjorkegren, and J. Tegner, “Consistent feature selection for pattern recognition in polynomial time,” *Journal of Machine Learning Research*, vol. 8, pp. 589–612, 2007.

[83] S. Baek, C. Tsai, and J. Chen, “Development of biomarker classifiers from high-dimensional data,” *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 537–546, 2009.